SVD and Least Squares Problems

Note Title

LECTURE 17

☆ LS via SVD

Recall the LS solution via QR factorization:

(1) Compute reduced QR of A.
(2) Compute $\mathbf{y} = \hat{Q}^T \mathbf{b}$.
(3) Solve $\hat{R}\mathbf{x} = \mathbf{y}$ — (*)

If A: full rank, then $\hat{R}_{ii} \neq 0$, $1 \le i \le n$, and the triangular system (*) has a unique LS solution.

Now using the reduced SVD of A, i.e., $A = \hat{U}\hat{\Sigma}V^T$, we can also solve the normal eqn:

$$A^T A \mathbf{x} = A^T \mathbf{b}$$

$$\Leftrightarrow \quad (\hat{U}\hat{\Sigma}V^T)^T(\hat{U}\hat{\Sigma}V^T)\mathbf{x} = (\hat{U}\hat{\Sigma}V^T)^T \mathbf{b}$$

$$\Leftrightarrow \quad V\hat{\Sigma}^T\hat{U}^T\hat{U}\hat{\Sigma}V^T\mathbf{x} = V\hat{\Sigma}\hat{U}^T\mathbf{b}$$

$$\Leftrightarrow \quad V\hat{\Sigma}^T\hat{\Sigma}V^T\mathbf{x} = V\hat{\Sigma}^T\hat{U}^T\mathbf{b}$$

$$\Leftrightarrow \quad \hat{\Sigma}^T\hat{\Sigma}V^T\mathbf{x} = \hat{\Sigma}^T\hat{U}^T\mathbf{b} \quad \text{since } V\text{: ortho.}$$

$$\Leftrightarrow \quad \hat{\Sigma}V^T\mathbf{x} = \hat{U}^T\mathbf{b} \quad \text{if } A\text{: full rank, i.e., } \sigma_j > 0,\ 1 \le j \le n$$

This can be solved easily.

(1) Compute reduced SVD of A.
(2) Compute $\mathbf{y} = \hat{U}^T \mathbf{b}$.
(3) Solve $\hat{\Sigma}\mathbf{w} = \mathbf{y}$. —— (**)
(4) Set $\mathbf{x} = V\mathbf{w}$.

Note: (**) is a diagonal system, easier to solve than (*)!!

☆ Pseudoinverse and SVD

Recall that if $A \in \mathbb{R}^{m\times n}$ is full rank,

$\underline{m > n}$: $A^+ = (A^T A)^{-1} A^T$

$\underline{m = n}$: $A^+ = A^{-1}$

$\underline{m < n}$: $A^+ = A^T (A A^T)^{-1}$

However, we can define the pseudo inv. using SVD even if A is not full rank!

$$A = U \Sigma V^T, \qquad \Sigma = \begin{array}{c} \begin{matrix} \overbrace{\qquad\qquad}^{r} & \overbrace{\quad}^{n-r} \end{matrix} \\ \left[\begin{array}{cc|c} \sigma_1 & 0 & 0 \\ 0 & \sigma_r & \\ \hline \multicolumn{2}{c|}{0} & 0 \end{array}\right] \begin{matrix} \}\, r \\ \\ \}\, m-r \end{matrix} \end{array}$$

Define

$$A^+ := V \Sigma^+ U^T, \qquad \Sigma^+ := \begin{array}{c} \begin{matrix} \overbrace{\qquad\qquad}^{r} & \overbrace{\quad}^{m-r} \end{matrix} \\ \left[\begin{array}{cc|c} 1/\sigma_1 & 0 & 0 \\ 0 & 1/\sigma_r & \\ \hline \multicolumn{2}{c|}{0} & 0 \end{array}\right] \begin{matrix} \}\, r \\ \\ \}\, n-r \end{matrix} \end{array}$$

As we discussed before, A^+ satisfies the following Moore-Penrose conditions:

(i) $AXA = A$; (ii) $XAX = X$

(iii) $(AX)^T = AX$; (iv) $(XA)^T = XA$.

Such X is uniquely determined and $X = A^+$!!

★ Pseudoinverse & Orthogonal Projectors

Thm AA^{+} is an ortho. proj. onto range(A)
and $AA^{+} = U_r U_r^T$
$A^{+}A$ is an ortho. proj. onto range(A^T)
and $A^{+}A = V_r V_r^T$
where $U_r \in \mathbb{R}^{m\times r}$, $V_r \in \mathbb{R}^{n\times r}$ consist of the first r columns of U, V, respectively. $r = \text{rank}(A)$.

(Proof) Let $P_A := AA^{+}$, $P_{A^T} := A^{+}A$.
Now, $P_A = U\Sigma V^T V \Sigma^{+} U^T$

$$= U\Sigma\Sigma^{+}U^T = U \left[\begin{array}{c|c} I_r & 0 \\ \hline 0 & 0 \end{array}\right] U^T$$
$$= U_r U_r^T \quad \checkmark$$

$$P_A^2 = U_r \underbrace{U_r^T U_r}_{= I_r} U_r^T = U_r U_r^T = P_A \quad \checkmark$$
so it's a proj.!

$$P_A^T = (U_r U_r^T)^T = (U_r^T)^T U_r^T = U_r U_r^T = P_A \quad \checkmark$$
So it's an ortho. proj.!

Finally, it's also clear that P_A maps onto range(A) since range$(A) = \langle u_1, \cdots, u_r \rangle$. $\checkmark$
You can do similarly for P_{A^T} ///

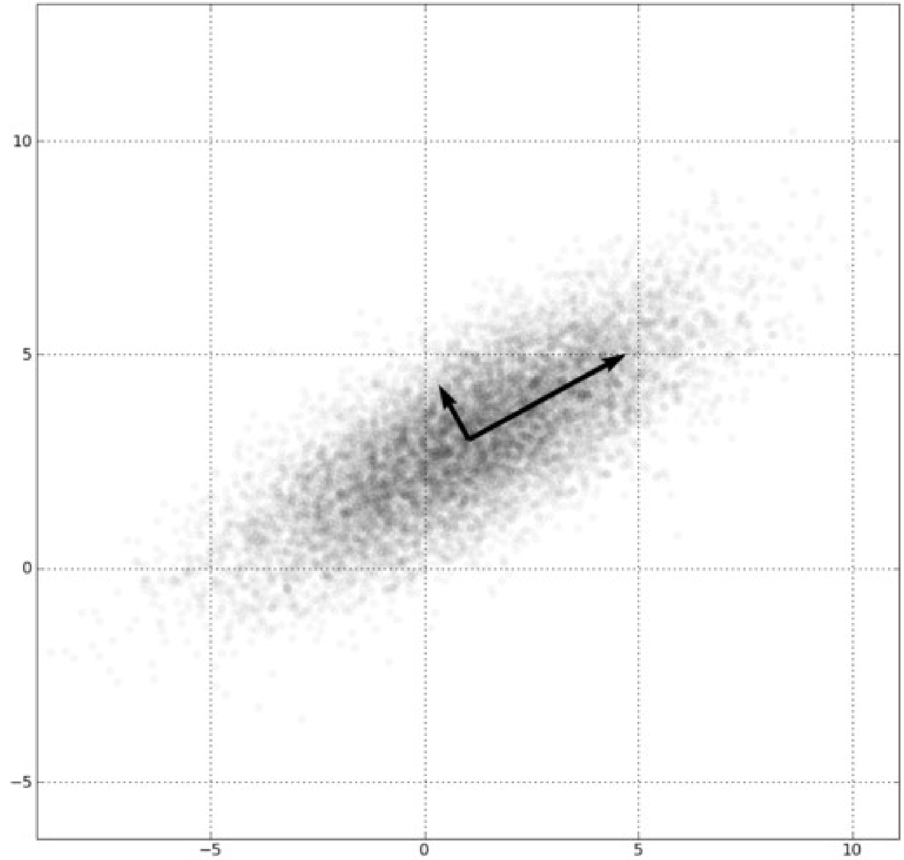
Note: Consider any $x \in$ range(A).
Then $\exists\, y \in \mathbb{R}^n$ s.t. $x = Ay$.
Now $P_A x = AA^{+}x = \underbrace{AA^{+}A}\, y$
$$= Ay = x.$$
"A via Moore-Penrose (i)

★ Principal Component Analysis (PCA)

(a.k.a. Karhunen-Loève Transform)

is a data analysis technique that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called "principal components."

2D example (from Wikipedia)



One can understand PCA using SVD! But before doing so, we need a bit of statistics.

often these are viewed as n realizations of some stochastic process.

Suppose we are given a set of vectors (observations)

$$\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$$

and each $\mathbf{x}_j \in \mathbb{R}^d$. d : could be huge (ex. a face image database).

Let $X := [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n] \in \mathbb{R}^{d \times n}$

You know the mean (or average) of this data set

$$\bar{\mathbf{x}} := \frac{1}{n} \sum_{j=1}^{n} \mathbf{x}_j$$

And define the centered data matrix

$$\tilde{X} := [\mathbf{x}_1 - \bar{\mathbf{x}} \ \ \mathbf{x}_2 - \bar{\mathbf{x}} \ \cdots \ \mathbf{x}_n - \bar{\mathbf{x}}]$$

Note : $\tilde{X} = X \left(I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T \right)$

↳ Good exercise!

Now the sample covariance matrix S is defined as

$$S := \frac{1}{n} \tilde{X} \tilde{X}^T \in \mathbb{R}^{d \times d}$$

S_{ij} indicates the covariance or mutual correlation between the ith and jth entries of data vectors.

PCA is nothing but an eigenvalue decomposition of S, i.e.,

$$S = \Phi \Lambda \Phi^T, \quad \Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_d)$$

Let's sort λ_i's as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$

Because $S^T = S$, and $S = \frac{1}{n}\tilde{X}\tilde{X}^T$, we can show that $\lambda_i \geq 0$. $1 \leq i \leq d$.

$\Phi = [\phi_1 \cdots \phi_d] \in \mathbb{R}^{d\times d}$ is a matrix containing the eigenvectors. also thanks to $S^T = S$, Φ is an orthogonal matrix whose columns form an ONB of $\mathbb{R}^d$.

The change of the bases from $[e_1 \cdots e_d]$ to $[\phi_1 \cdots \phi_d]$ is achieved simply by $\Phi^T \tilde{X}$.

$\phi_j^T \tilde{X}$ is called the j th principal components of X.

PCA was known for a long time, e.g., since the time of Pearson (1901) and Hotelling (1933).

Those days, the measurement dimension d was much smaller than the number of samples n, i.e. $d \ll n$

This is called the "classical" setting.

Ex. 5 exam scores of 2000 students $d=5$, $n=2000$.

Due to the advent of computers and sensor technology, now we often have $d \gg n$, the "neo-classical" setting.

Ex. The face database: $d = 128^2$, $n = 143$.